DOCUMENT RESUME

ED 408 305 TM 026 507

AUTHOR Kim, Seock-Ho

TITLE An Evaluation of Hierarchical Bayes Estimation for the Two-Parameter Logistic Model.

PUB DATE Mar 97

NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 1997).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Bayesian Statistics; Difficulty Level; *Estimation (Mathematics); *Item Bias; Maximum Likelihood Statistics; Sample Size; *Test Items

IDENTIFIERS *Hierarchical Analysis; Item Discrimination (Tests); Two Parameter Model

ABSTRACT

Hierarchical Bayes procedures for the two-parameter logistic item response model were compared for estimating item parameters. Simulated data sets were analyzed using two different Bayes estimation procedures, the two-stage hierarchical Bayes estimation (HB2) and the marginal Bayesian with known hyperparameters (MB), and marginal maximum likelihood estimation (ML). Three different prior distributions were employed in the two Bayes estimation procedures. HB2 and MB yielded consistently smaller root mean square differences and mean euclidean distances than ML. The HB2 and MB estimates of item discrimination parameters yielded relatively larger biases than the ML estimates. As the sample size increased, the three estimation procedures yielded essentially the same bias pattern for item discrimination. Bias results of item difficulty show no differences among the estimation procedures. Tight prior conditions yielded smaller root mean square differences and mean euclidean distances. An appendix discusses the estimate of the unknown item parameters in detail. (Contains 2 figures, 4 tables, and 45 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 408 305

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Seock-Ho Kim
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

An Evaluation of Hierarchical Bayes Estimation for the Two-Parameter Logistic Model

Seock-Ho Kim
The University of Georgia

March, 1997
Running Head: HIERARCHICAL BAYES ESTIMATION

Paper presented at the annual meeting of the American Educational Research Association, Chicago.

BEST COPY AVAILABLE

TM026507

An Evaluation of Hierarchical Bayes Estimation for the Two-Parameter Logistic Model

Abstract

Hierarchical Bayes procedures for the two-parameter logistic item response model were compared for estimating item parameters. Simulated data sets were analyzed using two different Bayes estimation procedures, the two-stage hierarchical Bayes estimation (HB2) and the marginal Bayesian with known hyperparameters (MB), and marginal maximum likelihood estimation (ML). Three different prior distributions were employed in the two Bayes estimation procedures. HB2 and MB yielded consistently smaller root mean square differences and mean euclidean distances than ML. The HB2 and MB estimates of item discrimination parameters yielded relatively larger biases than the ML estimates. As the sample size increased, the three estimation procedures yielded essentially the same bias pattern for item discrimination. Bias results of item difficulty show no differences among the estimation procedures. Tight prior conditions yielded smaller root mean square differences and mean euclidean distances.

Key words: Bayes estimation, hierarchical prior, item response theory, marginal Bayesian estimation, maximum likelihood estimation.

Introduction

Ever since Birnbaum (1969) presented Bayes methods of estimating ability parameters, a number of Bayesian approaches have been proposed under item response theory (IRT) for estimating item and ability parameters. The key feature of the Bayesian approach is its reliance upon simple probability theory that provides a theoretical framework for incorporating prior information or belief into the estimation of parameters to improve accuracy of estimates.

Currently the Bayesian approaches in IRT can be distinguished by whether the estimation of item parameters takes place with marginalization over incidental ability parameters (Mislevy, 1986; Tsutakawa & Lin, 1986) or without any marginalization (Kim, Cohen, Baker, Subkoviak, & Leonard, 1994; Swaminathan & Gifford, 1982, 1985, 1986). The marginal modes may provide better approximations to the posterior means in the presence of nuisance parameters than the joint modes (Mislevy, 1986; O'Hagan, 1976; Tsutakawa & Lin, 1986). This point has been empirically demonstrated by Kim et al. (1994), especially for small data sets.

Since specification of priors in Bayesian analysis is a subjective matter, a number of different forms of priors have been studies in estimation of item parameters. The hierarchical Bayes approach, suggested by Good (1980, 1983), Lindley (1971), and Lindley and Smith (1972), has been successfully applied to the estimation of item and ability parameters (Mislevy, 1986; Swaminathan & Gifford, 1982, 1985, 1986). Kim (1994) presented a two-stage hierarchical Bayes estimation of item parameters which involved in marginalization over incidental ability parameters (i.e., marginal Bayesian estimation with a two-stage hierarchical prior). Kim (1994) compared the item parameter estimates yielded by this two-stage hierarchical Bayes estimation with those obtained via maximum likelihood estimation and via other marginal Bayesian estimation procedures using LSAT-6 and LSAT-7 data sets (Bock & Lieberman, 1970). He found that the item parameter estimates yielded by the marginal Bayesian estimation procedures with different prior distributions were very similar. Parameter estimates yielded by the empirical Bayes estimation procedure for LSAT-6 and LSAT-7 were different from those yielded by other estimation procedures. However, these results were based on limited examples. It is of interest, thus, to compare the characteristics of item parameter estimates yielded by the two-stage hierarchical Bayes estimation with those obtained via marginal Bayesian estimation with different priors and via marginal maximum

likelihood estimation in a recovery study context.

Complete exploitation of the potential of the Bayesian estimation requires understanding of its mathematical underpinnings, particularly the role of prior distributions in estimating parameters. In a classical Bayesian approach, a single prior can be selected for the ordinary parameters. It is possible to recognize some uncertainty in priors. When priors are expressed in terms of family or class of prior, we call the parameters in the class of priors as hyperparameters. Hyperparameters describe the distributional characteristics of the prior information. It is sometimes also convenient to specify prior information on the hyperparameters as well. This second prior is called a hyperprior and contains parameters which are referred to as hyperhyperparameters (Good, 1980, 1983; Lindley, 1971, Lindley & Smith, 1972).

In this paper, we first present marginal Bayesian estimation of item parameters with a two-stage hierarchical prior distribution for the two-parameter logistic IRT model. Next, we present empirical comparisons among two Bayes estimation procedures (i.e., the two-stage hierarchical Bayes estimation and the marginal Bayesian estimation with known hyperparameters) and the marginal maximum likelihood estimation procedure. Three different priors were employed in the Bayes estimation procedures. It can be noted that point estimates of the ability parameters do not arise during the course of the marginal Bayesian estimation of item parameters. They are calculated after obtaining the estimates of item parameters, assuming the item parameters to be known (Bock & Aitkin, 1981; Mislevy & Bock, 1990). We do not discuss the estimation of ability parameters in this paper.

Theoretical Framework

IRT Model and Marginalization

Consider binary responses to a test with n items by each of N examinees. A response of examinee i to item j is represented by a random variable Y_{ij}, where $i = 1, \ldots, N$ and $j = 1, \ldots, n$. The probability of a correct response of examinee i to item j is represented by $P(Y_{ij} = 1|\theta_i, \boldsymbol{\xi}_j) = P_j(\theta_i)$ and the probability of an incorrect response is given by $P(Y_{ij} = 0|\theta_i, \boldsymbol{\xi}_j) = 1 - P_j(\theta_i) = Q_j(\theta_i)$, depending on a real-valued ability parameter θ_j and a real- or vector-valued item parameter $\boldsymbol{\xi}_j$. For the two-parameter logistic model, the probability of a correct response has the form

$$P_j(\theta_i) = \frac{1}{1 + \exp\{-a_j(\theta_i - b_j)\}}, \tag{1}$$

where $\boldsymbol{\xi}_j = (a_j, b_j)'$, and a_j and b_j are the item discrimination and difficulty parameters, respectively.

For examinee i, there is an observed vector of dichotomously scored item responses of length n denoted by $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})$. Under the assumption of conditional independence, the probability of $\mathbf{y}_i$ given θ_i and the vector of all item parameters, $\boldsymbol{\xi} = (a_1, b_1, \ldots, a_n, b_n)'$, is

$$p(\mathbf{y}_i|\theta_i, \boldsymbol{\xi}) = \prod_{j=1}^{n} P_j(\theta_i)^{y_{ij}} Q_j(\theta_i)^{1-y_{ij}}. \tag{2}$$

The marginal probability of obtaining the response vector $\mathbf{y}_i$ for examinee i sampled from a given population is

$$p(\mathbf{y}_i|\boldsymbol{\xi}) = \int_{\Theta} p(\mathbf{y}_i|\theta_i, \boldsymbol{\xi}) p(\theta_i) d\theta_i, \tag{3}$$

where Θ is the parameter space and $p(\theta_i)$ is a continuous population distribution of θ_i. Without loss of generality, we assume that θ_i are independent and identically distributed as standard normal, $\theta_i \sim N(0, 1)$. This assumption can be relaxed as the ability distribution may be empirically characterized (Bock & Aitkin, 1981). The marginal probability of $\mathbf{y}_i$ can be approximated with any specified degree of precision by Gaussian quadrature formulas (Stroud & Secrest, 1966) using

$$p(\mathbf{y}_i|\boldsymbol{\xi}) = \sum_{k=1}^{q} p(\mathbf{y}_i|X_k, \boldsymbol{\xi}) A(X_k), \tag{4}$$

where X_k are called the nodes and $A(X_k)$ are the corresponding weights. Since we assume θ_i are randomly sampled from $N(0, 1)$, we may use Gauss-Hermite quadratures, for example, $X_k = \sqrt{2} X_k^*$ and $A(X_k) = A(X_k^*)/\sqrt{\pi}$, where X_k^* and $A(X_k^*)$ are obtained from Stroud and Secrest (1966).

The marginal probability of obtaining the $N \times n$ response matrix $\mathbf{y}$ is then given by

$$p(\mathbf{y}|\boldsymbol{\xi}) = \prod_{i=1}^{N} p(\mathbf{y}_i|\boldsymbol{\xi}) = l(\boldsymbol{\xi}|\mathbf{y}), \tag{5}$$

where $l(\boldsymbol{\xi}|\mathbf{y})$ may be regarded as a function of $\boldsymbol{\xi}$ given the data $\mathbf{y}$. Bayes' theorem tells us that the posterior probability distribution for $\boldsymbol{\xi}$ to the data $\mathbf{y}$ is proportional to the product of the likelihood for $\boldsymbol{\xi}$ given $\mathbf{y}$ and the distribution for $\boldsymbol{\xi}$ prior to the data. That is,

$$p(\boldsymbol{\xi}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\xi}) p(\boldsymbol{\xi})}{p(\mathbf{y})} \propto l(\boldsymbol{\xi}|\mathbf{y}) p(\boldsymbol{\xi}), \tag{6}$$

where $\propto$ denotes proportionality. The likelihood function represents the information about $\boldsymbol{\xi}$ obtained from the data through which the data $\mathbf{y}$ may modify our prior knowledge of $\boldsymbol{\xi}$. A

prior distribution represents what is known about unknown parameters before the data are obtained. Prior knowledge or relative ignorance can be represented by such a distribution.

Parameter Estimation in IRT

Lord (1986) presented advantages and disadvantages of several parameter estimation methods in IRT. Birnbaum (1968) and Lord (1980) recommend the estimation of the $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ by joint maximization of their likelihood function

$$p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\xi}) = \prod_{i=1}^{N}\prod_{j=1}^{n} P_j(\theta_i)^{y_{ij}} Q_j(\theta_i)^{1-y_{ij}} = l(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y}), \tag{7}$$

where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_N)'$. Especially, the item parameter estimation part for maximizing $l(\boldsymbol{\xi}|\mathbf{y},\hat{\boldsymbol{\theta}})$ and the ability parameter estimation part for maximizing $l(\boldsymbol{\theta}|\mathbf{y},\hat{\boldsymbol{\xi}})$ are iterated to obtain stable estimates of item and ability parameters.

Extending the idea of joint maximization, Swaminathan and Gifford (1982, 1985, 1986) suggested that $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ can be estimated by joint maximization with respect to these parameters of the posterior density

$$p(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y}) \propto l(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})p(\boldsymbol{\theta},\boldsymbol{\xi}), \tag{8}$$

where $p(\boldsymbol{\theta},\boldsymbol{\xi})$ is the joint prior density of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. On the assumption that priors of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are independently distributed with probability density functions $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\xi})$, the item parameter estimation part which maximizes $l(\boldsymbol{\xi}|\mathbf{y},\hat{\boldsymbol{\theta}})p(\boldsymbol{\xi})$ and the ability parameter estimation part which maximizes $l(\boldsymbol{\theta}|\mathbf{y},\hat{\boldsymbol{\xi}})p(\boldsymbol{\theta})$ are iterated to obtain stable estimates of item and ability parameters.

Alternatively, Bock and Aitken (1981), Bock and Lieberman (1970), Harwell, Baker, and Zwarts (1988), and Tsutakawa (1984) presented estimation of $\boldsymbol{\xi}$ by maximization of the marginal or integrated likelihood in Equation 5. The development of marginal maximum likelihood estimation was motivated by the structural and incidental parameters problem (Baker, 1987). Assuming that the IRT model and the ability distribution are properly specified, the resulting item parameter estimates are consistent for tests of finite length (Bock & Aitkin, 1981).

Since the marginal likelihood in Equation 5 is not a probability density function, we cannot make a probabilistic statement regarding $\boldsymbol{\xi}$. We can accomplish this by analyzing the marginal posterior distribution in Equation 6 (e.g., Harwell & Baker, 1991; Leonard

& Novick, 1985; Mislevy, 1986; Tsutakawa, 1992; Tsutakawa & Lin, 1986). The posterior density represents a compromise between the likelihood and the prior density. Hence, an important element of Bayesian inference is the prior information concerning $\boldsymbol{\xi}$. In Bayesian analysis it is necessary to have a convenient way to quantify such information.

Prior and Posterior Distribution

Prior information for parameters is expressed in terms of probability distributions in the Bayesian approach. It can be noted that a flexible family of prior distributions is available by transforming item parameters into new parameters which may be distributed as a multivariate normal distribution. Following Leonard and Novick (1985) and Mislevy (1986), we use the transformation $\alpha_j = \log a_j$. We may also write $\beta_j = b_j$ and $\boldsymbol{\xi}_j = (\alpha_j, \beta_j)'$.

We assume that the vector of item parameters $\boldsymbol{\xi}$ possesses a multivariate normal distribution conditional on the respective mean vector $\boldsymbol{\mu}_\xi$ and covariance matrix $\boldsymbol{\Sigma}_\xi$. The complete form of the hierarchical prior distribution of item parameters is given by

$$p(\boldsymbol{\xi}, \boldsymbol{\eta}) = p_1(\boldsymbol{\xi}|\boldsymbol{\eta}) p_2(\boldsymbol{\eta}), \tag{9}$$

where the hyperparameter $\boldsymbol{\eta} = (\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$, and the subscripts 1 and 2 denote the first stage and the second stage, respectively, of the prior distribution.

If we assume the vectors of item parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are independent, we can take the vectors to possess independent multivariate normal distributions, conditional on their mean vectors, $\boldsymbol{\mu}_\alpha$ and $\boldsymbol{\mu}_\beta$, and covariance matrices, $\boldsymbol{\Sigma}_\alpha$ and $\boldsymbol{\Sigma}_\beta$. Then

$$p_1(\boldsymbol{\xi}|\boldsymbol{\eta}) p_2(\boldsymbol{\eta}) = p_1(\boldsymbol{\alpha}|\boldsymbol{\eta}_\alpha) p_1(\boldsymbol{\beta}|\boldsymbol{\eta}_\beta) p_2(\boldsymbol{\eta}_\alpha) p_2(\boldsymbol{\eta}_\beta), \tag{10}$$

where $\boldsymbol{\eta}_\alpha = (\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$ and $\boldsymbol{\eta}_\beta = (\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$.

When we further assume exchangeability for all parameters, we may take $\boldsymbol{\mu}_\alpha = \mu_\alpha \mathbf{1}$, $\boldsymbol{\Sigma}_\alpha = \sigma_\alpha^2 \mathbf{I}_n$, $\boldsymbol{\mu}_\beta = \mu_\beta \mathbf{1}$, and $\boldsymbol{\Sigma}_\beta = \sigma_\beta^2 \mathbf{I}_n$, where μ_α, σ_α^2, μ_β, and σ_β^2 are scalars, $\mathbf{1}$ is an $n \times 1$ vector of ones, and $\mathbf{I}_n$ is an identity matrix of order n (Leonard & Novick, 1985). The first stage prior distribution can be expressed as

$$p_1(\boldsymbol{\xi}|\boldsymbol{\eta}) = \prod_{j=1}^{n} p_1(\alpha_j|\mu_\alpha, \sigma_\alpha^2) p_1(\beta_j|\mu_\beta, \sigma_\beta^2), \tag{11}$$

where

$$p_1(\alpha_j|\mu_\alpha, \sigma_\alpha^2) = (2\pi\sigma_\alpha^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma_\alpha^2}(\alpha_j - \mu_\alpha)^2\right\}, \tag{12}$$

and $p_1(\beta_j|\mu_\beta, \sigma_\beta^2)$ can be similarly defined. A hierarchical Bayes approach then assigns another stage priors to the hyperparameter $\boldsymbol{\eta}$.

Hyperpriors for μ_α and σ_α^2 can be specified by assuming that μ_α has a noninformative uniform distribution and $\nu_\alpha\lambda_\alpha/\sigma_\alpha^2$ is distributed as $\chi_{\nu_\alpha}^2$, where ν_α is the degrees of freedom. Hence,

$$p_2(\boldsymbol{\eta}_\alpha) = p_2(\mu_\alpha)p_2(\sigma_\alpha^2|\nu_\alpha, \lambda_\alpha) = \frac{(\sigma_\alpha^2)^{-(\nu_\alpha/2+1)}}{\Gamma(\nu_\alpha/2)(\nu_\alpha\lambda_\alpha/2)^{-\nu_\alpha/2}} \exp\left(-\frac{\nu_\alpha\lambda_\alpha}{2\sigma_\alpha^2}\right), \tag{13}$$

where ν_α and λ_α are hyperhyperparameters. Now the prior distribution for $\boldsymbol{\alpha}$ can be expressed as

$$p_1(\boldsymbol{\alpha}|\boldsymbol{\eta}_\alpha)p_2(\boldsymbol{\eta}_\alpha) =$$
$$(2\pi\sigma_\alpha^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_\alpha^2}\sum_{j=1}^{n}(\alpha_j - \mu_\alpha)^2\right\} \times \frac{(\sigma_\alpha^2)^{-(\nu_\alpha/2+1)}}{\Gamma(\nu_\alpha/2)(\nu_\alpha\lambda_\alpha/2)^{-\nu_\alpha/2}} \exp\left(-\frac{\nu_\alpha\lambda_\alpha}{2\sigma_\alpha^2}\right). \tag{14}$$

The above equation depends upon nuisance parameters, μ_α and σ_α^2, and these can be integrated out. Integrating out μ_α and σ_α^2 yields

$$\int_0^\infty \int_{-\infty}^\infty p_1(\boldsymbol{\alpha}|\boldsymbol{\eta}_\alpha)p_2(\boldsymbol{\eta}_\alpha)d\mu_\alpha d\sigma_\alpha^2 = p(\boldsymbol{\alpha}|\nu_\alpha, \lambda_\alpha) \propto \left\{\sum_{j=1}^{n}(\alpha_j - \bar{\alpha})^2 + \nu_\alpha\lambda_\alpha\right\}^{-(n+\nu_\alpha-1)/2}. \tag{15}$$

Similar specification yields $p(\boldsymbol{\beta}|\nu_\beta, \lambda_\beta)$. As we integrated out the hyperparameter $\boldsymbol{\eta}$, we can express the prior distribution of item parameters as

$$p(\boldsymbol{\xi}|\boldsymbol{\eta}^{(2)}) = p(\boldsymbol{\alpha}|\nu_\alpha, \lambda_\alpha)p(\boldsymbol{\beta}|\nu_\beta, \lambda_\beta), \tag{16}$$

where the hyperhyperparameter $\boldsymbol{\eta}^{(2)} = (\nu_\alpha, \lambda_\alpha, \nu_\beta, \lambda_\beta)$.

In fact, the complete prior for the hierarchical model, assuming independence between ability and item parameters, can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\eta}) = p(\boldsymbol{\theta}, \boldsymbol{\tau})p(\boldsymbol{\xi}, \boldsymbol{\eta}) = p_1(\boldsymbol{\theta}|\boldsymbol{\tau})p_2(\boldsymbol{\tau})p_1(\boldsymbol{\xi}|\boldsymbol{\eta})p_2(\boldsymbol{\eta}), \tag{17}$$

where $p_1(\boldsymbol{\theta}|\boldsymbol{\tau})$ is the first stage density of $\boldsymbol{\theta}$ conditional on $\boldsymbol{\tau}$, $\boldsymbol{\tau}$ are examinee population parameters which takes the second stage density $p_2(\boldsymbol{\tau})$, $p_1(\boldsymbol{\xi}|\boldsymbol{\eta})$ is the first stage density of $\boldsymbol{\xi}$ conditional on $\boldsymbol{\eta}$, and $\boldsymbol{\eta}$ are item population parameters which follows the second stage density $p_2(\boldsymbol{\eta})$ (Mislevy, 1986). In this paper, we assumed $\boldsymbol{\tau} = (\mu_\theta, \sigma_\theta^2) = (0, 1)$ is given and $\boldsymbol{\eta} = (\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$ is integrated out. The prior distribution of both item and ability parameters given $\boldsymbol{\tau}$ and the hyperhyperparameter $\boldsymbol{\eta}^{(2)}$ can be written as

$$p(\boldsymbol{\theta}|\boldsymbol{\tau})p(\boldsymbol{\xi}|\boldsymbol{\eta}^{(2)}). \tag{18}$$

The marginal posterior distribution given $\boldsymbol{\tau}$ and $\boldsymbol{\eta}^{(2)}$ is then

$$p(\boldsymbol{\xi}|\mathbf{y}, \boldsymbol{\tau}, \boldsymbol{\eta}^{(2)}) \propto p(\mathbf{y}|\boldsymbol{\xi}, \boldsymbol{\tau}) p(\boldsymbol{\xi}|\boldsymbol{\eta}^{(2)}) = l(\boldsymbol{\xi}|\mathbf{y}, \boldsymbol{\tau}) p(\boldsymbol{\xi}|\boldsymbol{\eta}^{(2)}). \tag{19}$$

Marginal Bayesian modal estimates of item parameters can be found by maximizing the marginal posterior distribution with respect to $\boldsymbol{\xi}$. Appendix presents a brief description of procedures for implementation of the marginal Bayes modal estimation with the two-stage hierarchical priors.

Method

Data were simulated under the following conditions: (1) number of examinees (N = 100, 300), (2) number of items (n = 15, 45), (3) estimation (HB2, MB, ML), and (4) prior condition (prior-α_{L}, prior-α_{T}, prior-$\alpha\beta_{\mathrm{T}}$). The sample sizes and the test lengths were selected to emulate the situation in which estimation procedures and priors might have some impact upon item and ability parameter estimates. The sample size and the test length were completely crossed to yield four situations.

Three estimation procedures were used; the two-stage hierarchical Bayes estimation (HB2), the marginal Bayesian with known hyperparameters (MB), and marginal maximum likelihood estimation (ML). The two Bayes estimation procedures, HB2 and MB, had the three prior conditions: prior-α_{L}, prior-α_{T}, and prior-$\alpha\beta_{\mathrm{T}}$. The prior-α_{L} condition used a loose prior for the transformed item discrimination; the prior-α_{T} condition used a tight prior for the transformed item discrimination; and the prior-$\alpha\beta_{\mathrm{T}}$ condition used tight priors for both the transformed item discrimination and the item difficulty. The exact specification of each prior condition is presented in a subsequent section on the item and ability parameter estimation. ML, of course, did not employ a prior distribution in estimation.

Data Generation

The data sets used in this study were the same as those used in Kim et al. (1994). Dichotomous item response vectors were generated using the two-parameter logistic model via the computer program GENIRV (Baker, 1982). Based on the usual ranges of item parameters for the two-parameter logistic model, the underlying transformed item discrimination parameters were assumed to be normally distributed with mean 0 and variance .09, $\alpha_j \sim N(0, .09)$. The underlying item discrimination parameters a_j are distributed with

mean 1.046 and variance .103. The underlying item difficulty parameters are distributed normally with mean 0 and variance 1, $b_j \sim N(0,1)$. For data generation purposes, an approximation based on histograms was adopted instead of selecting item parameters randomly from a specified distribution. Item discrimination and item difficulty parameters for the 15-item test were set to have three different values (the number of items is given in parentheses): Item discrimination parameters were .66 (4), 1 (7), and 1.51 (4), and item difficulty parameters were −1.38 (4), 0 (7), and 1.38 (4). For the 45-item test, each of the item parameters was set to have five different values: Item discrimination parameters were .57 (4), .76 (9), 1 (19), 1.32 (9), and 1.77 (4), and item difficulty parameters were −1.9 (4), −.95 (9), 0 (19), .95 (9), and 1.9 (4). There was no correlation between item discrimination and difficulty parameters.

The underlying ability parameters were matched to the item difficulty distribution. Hence, a normal distribution with mean 0 and variance 1, $\theta_i \sim N(0,1)$, was used to specify the underlying ability parameters. Also, an approximation based on histograms was adopted for ability and yielded 11 ability levels. For the 100-examinee sample, the ability parameter were set to be −2.5 (1), −2 (3), −1.5 (7), −1 (12), −.5 (17), 0 (20), .5 (17), 1 (12), 1.5 (7), 2 (3), and 2.5 (1), where parentheses contain the number of examinees. For 300-examinee sample, the ability parameters were set to be −2.5 (4), −2 (8), −1.5 (20), −1 (36), −.5 (52), 0 (60), .5 (52), 1 (36), 1.5 (20), 2 (8), and 2.5 (4).

For each of the factors of sample size and test length, four replications of the simulated data were generated. Since the two factors were completely crossed, a total of 16 GENIRV runs was needed to obtain the data sets for the study.

Item and Ability Parameter Estimation

Each of the generated data sets was analyzed via the computer program BILOG (Mislevy & Bock, 1990) for the MB and ML procedures and via the computer program HBAYES, specifically developed for this study to provide the HB2 estimates. In each Bayes estimation procedure, three prior conditions, prior-α_L, prior-α_T, and prior-$\alpha\beta_T$, were employed. Note that a prior was not employed in ML. Hence, for example, the generated item response data set for the first replication of sample size 100 and test length 15 was analyzed by seven computer runs (two Bayes estimation procedures with three prior conditions and maximum likelihood estimation).

In the prior-α_L condition for MB, a lognormal prior with mean 0 and variance .25 was

used, that is, $\ln a_j \sim N(0, .25)$. This is, in fact, the default prior specification in BILOG for the estimation of item parameters of the two-parameter logistic model. In the prior-α_T condition for MB, a lognormal distribution with mean 0 and variance .09, $\ln a_j \sim N(0, .09)$, was used. For the prior-$\alpha\beta_T$ condition for MB, the same prior in the prior-α_T condition along with a normal prior was used for the item difficulty with mean 0 and variance 1, $\beta_j \sim N(0, 1)$.

For HB2, the mean hyperparameter was assumed to have a noninformative uniform distribution and the variance hyperparameter was set to have an inverse chi-square distribution. In the prior-α_L condition, the inverse chi-square distribution with $\nu_\alpha = 8$ and $\lambda_\alpha = .25$ was used for the variance hyperparameter of the transformed item discrimination parameters: $\nu_\alpha \lambda_\alpha / \sigma_\alpha^2 \sim \chi^2_{\nu_\alpha}$ and, thus, $2/\sigma_\alpha^2 \sim \chi^2_8$. The inverse chi-square distribution with parameters $\nu_\alpha = 8$ and $\lambda_\alpha = .09$ was used in the prior-α_T condition: $.72/\sigma_\alpha^2 \sim \chi^2_8$. Two inverse chi-square distributions with parameters $\nu_\alpha = 8$ and $\lambda_\alpha = .09$, and $\nu_\beta = 8$ and $\lambda_\beta = 1$ for the variance hyperparameters of the transformed item discrimination and of the item difficulty, respectively, were adopted for the prior-$\alpha\beta_T$ condition: $.72/\sigma_\alpha^2 \sim \chi^2_8$ and $8/\sigma_\beta^2 \sim \chi^2_8$.

When the mean hyperparameter is assumed to have a fixed value, μ, the specification of the variance hyperparameter by the inverse chi-square distribution with parameters ν and λ (i.e., $\nu\lambda/\sigma^2 \sim \chi^2_\nu$) yields the parameter of interest which is distributed as a t with mean μ, variance λ, and degrees of freedom ν, $t(\nu, \mu, \lambda)$ (Berger, 1985). Therefore, for the transformed item discrimination, assuming the mean hyperparameter μ_α has a fixed value, specification of the hyperparameter of variance by the inverse chi-square with $\nu_\alpha = 8$ and $\lambda_\alpha = .25$ yields a transformed item discrimination parameter which is distributed as a t with mean μ_α, variance $\lambda_\alpha = .25$, and degrees of freedom $\nu_\alpha = 8$, that is, $\alpha_j \sim t(8, \mu_\alpha, .25)$. Similarly, the specification with $\nu_\alpha = 8$ and $\lambda_\alpha = .09$ implies $\alpha_j \sim t(8, \mu_\alpha, .09)$; and the specification with $\nu_\beta = 8$ and $\lambda_\beta = 1$ yields $\beta_j \sim t(8, \mu_\beta, 1)$. In the above illustration, because we assumed a noninformative prior for the mean hyperparameter, the specifications used in HB2 will not produce the same specifications of item hyperparameters used in MB. These specifications are similar to their counterparts in MB.

Metric Transformation

In parameter recovery studies, such as the present one, comparisons between two or more sets of estimates and the underlying parameters require that the item and ability estimates

obtained from different calibration runs and their parameters be placed on a common metric (Baker & Al-Karni, 1991; Yen, 1987). Parameter estimation procedures under IRT yield metrics which are unique up to a linear transformation. To link both sets of estimates and parameters, it is necessary to determine the slope and intercept of the equating coefficients required for the transformation. The estimates of the item and ability parameters for each of the estimation procedures were placed on the scale of the true parameters using the test characteristic curve method by Stocking and Lord (1983) as implemented in the computer program EQUATE (Baker, 1993).

Criteria

The empirical evaluation in this study involved four criteria: root mean square difference (RMSD), correlation, and bias, and mean euclidean distance (MED). RMSD is the square root of the average of the squared differences between estimated and true values. For item discrimination, for example, RMSD is $\left\{(1/n)\sum_{j=1}^{n}(\hat{a}_j - a_j)^2\right\}^{1/2}$.

The bias B of a point estimator is the difference between the expected value of the estimates and the corresponding parameter (Mendenhall, Scheaffer, & Wackerly, 1981). The bias of the item discrimination estimates, for example, is given by $B_{a_j} = E(\hat{a}_j) - a_j$. The bias was obtained with regard to the underlying parameters across the four replications.

Since it is possible that an estimation procedure may function better at recovery of one type of item parameter than at recovery of the other, it is also useful to consider a single index which can describe simultaneously the quality of the recovery for both item parameters. MED provides such an index (Rudin, 1976). MED is the average of the square roots of the sum of the squared differences between the discrimination and difficulty parameter estimates and their generating values. MED is defined as $(1/n)\sum_{j=1}^{n}\left\{(\hat{\boldsymbol{\xi}}_j - \boldsymbol{\xi}_j)'(\hat{\boldsymbol{\xi}}_j - \boldsymbol{\xi}_j)\right\}^{1/2}$, where $\hat{\boldsymbol{\xi}}_j = (\hat{a}_j, \hat{b}_j)'$ and $\boldsymbol{\xi}_j = (a_j, b_j)'$. One caveat in using MED, of course, is that item discrimination and difficulty parameters are not expressed in comparable and interchangeable metrics. Even so, MED does provide a potentially useful descriptive index.

Results

RMSD and Correlation Results

Item Discrimination. Average RMSDs of item discriminations over four replications are reported in Table 1. As sample size increased, RMSDs decreased; marginal RMSD means were .265 and .185 for sample sizes 100 and 300, respectively.

Insert Table 1 about here

Two Bayes procedures, HB2 and MB, yielded smaller RMSDs than the ML procedure. For sample size 100, increasing the number of items increased the values of RMSD for HB2 but reduced the values of RMSD for ML. Increasing the number of items reduced the size of RMSDs for sample size 300. When the loose prior, α_L, was used in HB2 for sample size 100, it yielded comparatively smaller values of RMSD than did either of the tight prior conditions. The tight prior condition, α_T, in MB yielded smaller values of RMSD. There seem to be no differences among RMSDs when the tight prior conditions, α_T and $\alpha\beta_T$, were used for sample size 300.

The average correlations between true and estimated values of item discriminations across four replications are also given in Table 1. For each data set, HB2 yielded a slightly higher correlation than MB and ML. Generally, the larger the sample size, the higher the correlation. Also, increasing the number of items tended to produce higher correlations. For the three prior conditions used, no definitive tendency was observed in the correlations.

Item Difficulty. Table 2 contains the average RMSDs for item difficulty over four replications. An increase in sample size appeared to be associated with a decrease in the size of RMSDs. For sample size 100, increasing the number of items appeared to slightly decrease RMSDs except for ML. For sample size 300, increasing the number of items 15 to 45 resulted in larger values of RMSD. The values of RMSD from ML were consistently larger than the values from HB2 and MB regardless of sample sizes or test lengths.

Insert Table 2 about here

Prior-$\alpha\beta_T$ condition yielded a relatively smaller RMSDs than did either prior-α_L or prior-α_T conditions. HB2 consistently yielded smaller RMSDs than MB across all the prior conditions employed.

For each data set, all estimation procedures yielded nearly the same correlations between estimates and parameters (see Table 2). Generally, the larger sample sizes yielded higher correlations. Increasing the number of items yielded slightly higher correlations for 100-examinee data sets. This tendency was not observed for 300-examinee data sets. There

seemed to be no definitive trends in the correlations among the three prior conditions. ML yielded consistently lower correlations than did either HB2 or MB. It can be noticed, however, that all correlations were very high and close to 1.

Bias Results

Item Discrimination. The bias results for item discrimination, presented in Figure 1, appear to reflect influence by a number of factors. Each bias statistic was obtained by combining results from all four replications together.

Insert Figure 1 about here

For each test length, increasing sample size resulted in a decrease in bias values. In general, when Bayes estimation procedures were used, positive bias values were observed for the smaller item discrimination parameters (i.e., $a_j = .66$ for the 15-item test, and $a_j = .57$ and .76 for the 45-item test) due to regression toward the mean of the prior distribution. Conversely, negative values of bias were obtained for the relatively larger item discrimination parameters (i.e., $a_j = 1.51$ for the 15-item test, and $a_j = 1.32$ and 1.77 for the 45-item test). This shrinkage effect can be observed for nearly all data sets for HB2 and MB. HB2 yielded slightly more biased results.

Both tight prior conditions yielded relatively more biased results. The patterns of bias from HB2 and MB were very similar. ML yielded different patterns of bias than did the HB2 and MB procedures. The differences in bias patterns between ML and the two Bayes procedures were very pronounced in sample size 100. The differences diminished as the sample size increased to 300.

Item Difficulty. The bias results for item difficulty are reported in Figure 2. The pattern of results was somewhat different from that for item discrimination. For the 15-item test, all estimation procedures yielded nearly the same pattern of no bias. For the 45-item test, the three estimation procedures also resulted in nearly the same pattern of no bias for HB2, MB, and ML. For sample size 300, the patterns show nearly no bias results. Sample size 300 yielded relatively more stable bias results than sample size 100.

Insert Figure 2 about here

MED Results

Average MEDs between item parameter estimates and underlying item parameters over four replications are reported in Table 3. HB2 and MB yielded smaller MEDs than ML. For sample size 300, HB2 yielded smaller MEDs than MB under prior-α_L whereas HB2 yielded larger MEDs than MB under both prior-α_T and prior-$\alpha\beta_T$. For sample size 300, HB2 yielded consistently smaller MEDs than MB for all prior conditions. Also for sample size 300, the tight priors condition, $\alpha\beta_T$, yielded relatively smaller MEDs within each Bayes estimation. It can be noticed from Table 3 that MEDs decreased as the sample size increased. Increasing the number of items reduced the sizes of MED.

Insert Table 3 about here

Discussion

Maximum likelihood approaches in IRT suffer from a number of problems, an important one for the two-parameter logistic model being the possibility that outward drift of item discrimination estimates occurs and, consequently, unreasonable values will be obtained for parameter estimates. In addition, these approaches perform poorly when estimating item and ability parameters for unusual response patterns such as all correct or all incorrect answers. These problems have led to interest in the development of Bayesian approaches for estimation of item and ability parameters (Baker, 1987). In the present study, we used a recovery study approach to compare parameter estimates for the two-parameter logistic IRT model obtained via the two marginal Bayesian algorithms, HB2 and MB, and the maximum likelihood algorithm, ML.

Analysis of item parameter recovery results indicated that HB2 and MB yielded parameter estimates which were generally better than those obtained from ML. RMSD and MED results for item discrimination and difficulty were consistently larger for the ML estimates than for the HB2 and MB estimates. HB2 and MB estimates were similar although HB2 results were slightly better for prior-α_L and MB results were better for the tight prior conditions.

When $N = 300$, there seems to be essentially no bias in item discrimination estimates yielded by either HB2 or MB. Note that under ML, except for $N = 100$ and $n = 45$ data sets, there found positive values of bias for large values of item discrimination (1.51 when

n = 15 and 1.77 when n = 45). It should be noted that no incidence of nonconvergence due to outward drift of discrimination estimates occurred in ML for entire data sets. The bias results from item difficulty were almost identical for all estimation procedures and prior conditions. ML of course did not employ a prior distribution. All prior conditions contain a prior for item discrimination. Only prior-$\alpha\beta_T$ contained an additional prior for item difficulty. Many recovery studies indicate relatively excellent recovery results of item difficulty parameters. This might be a possible explanation why we have the same pattern of bias of item difficulty regardless of estimation procedures.

Both the shape and the variance of the prior distribution play a part in the Bayesian estimation of parameters. The more informative the prior, that is, the smaller the variance, the more the parameter estimate tends to be pulled toward the mean of the prior. In general, the use of tight priors seems appropriate when there is strong a priori information about the parameters. In the MB context, the same prior distributions were directly imposed on item parameters. Without the use of the empirical Bayes (i.e., FLOAT) option, the incorrect specification of the prior may result in more serious consequences for MB than HB2. Mislevy and Stocking (1989) recommended the use of the FLOAT option in BILOG when there is a possibility of mismatch between the expected value of item parameters and the prior mean. This issue was not tested in the present study because priors were relatively well matched to the generated data sets. In this regard, several issues remain to be studied in the present context. In particular, except Gifford and Swaminathan (1990) and Harwell and Janosky (1991), little has been done on the shrinkage effect. Neither are the effects of priors well known with respect to the robustness of the two-stage hierarchical model or other Bayes procedures. This kind of research is particularly valuable for small samples and short tests.

A prior distribution represents what is known about the parameter before the data are obtained. Consequently the role of the prior distribution is central in Bayesian analysis. The prior used in the Bayes procedures in this paper assumes independence and exchangeability among all item parameters. Sometimes dependence between item parameters should be considered. In this regard, Mislevy (1986) presented multivariate normal priors to account for dependency within item parameters. In addition, if the exchangeability of items cannot be exercised, we cannot use the same prior distribution for each item. Assuming all item parameter estimates and the corresponding estimated variance and covariance matrices from previous and possibly different calibrations were placed on the same ability metric, for

example, on the usual ability $N(0,1)$ metric, we can employ a different prior distribution for each item based upon existing information regarding the underlying item parameters.

In a usual Bayesian approach, prior distributions are used for the ordinary or transformed item parameters. To represent prior information of item parameters in terms of the item response function, confidence ellipsoids suggested by Thissen and Wainer (1990) can be helpful. In the two-stage hierarchical approach, we specified prior information on the hyperparameters. An alternative approach is to use a prior distribution based on entire item response function rather than item parameters. Tsutakawa and Lin (1986) suggested the use of an ordered bivariate beta prior distribution for values of the item response function at two ability levels. Also Tsutakawa (1992) suggested the use of the ordered Dirichlet prior on the entire item response function.

Note that the posterior density of hyperparameters may be closely approximated. For example, let $\mathbf{R}$ denote the $2n \times 2n$ posterior information matrix ($\mathbf{R} = -\mathbf{H}$), consisting of appropriate second derivatives of $\log p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{y})$ and evaluated at the marginal modes. Then the dispersion matrix $\mathbf{R}^{-1}$ provides an approximation to the posterior covariance matrix. In stead of estimating the prior parameters μ_α, σ_α^2, μ_β, and σ_β^2, following Leonard (1982), Leonard, Hsu, and Tsui (1989), and Tierney and Kadane (1986), we can approximate the posterior density of these hyperparameters by the Laplacian approximation

$$p(\mu_\alpha, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2|\mathbf{y}) \propto p(\mu_\alpha, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2) p(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}|\mathbf{y}) / |\mathbf{R}|^{1/2}. \tag{20}$$

Two possible choices of distribution for the prior parameters are the uniform distribution $p(\mu_\alpha, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2) \propto 1$, and the choice from Lindley and Smith (1972) which is

$$p(\mu_\alpha, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2) \propto \left(\sigma_\alpha^2\right)^{-(\nu_\alpha/2+1)} \left(\sigma_\beta^2\right)^{-(\nu_\beta/2+1)} \exp\left\{-\nu_\alpha \lambda_\alpha / 2\sigma_\alpha^2 - \nu_\beta \lambda_\beta / 2\sigma_\beta^2\right\}, \tag{21}$$

which takes all four hyperhyperparameters are independent, μ_α and μ_β each to be uniformly distributed over $(-\infty, \infty)$, and $\nu_\alpha \lambda_\alpha / \sigma_\alpha^2$ and $\nu_\beta \lambda_\beta / \sigma_\beta^2$ to possess chi-square distribution with respective degrees of freedom ν_α and ν_β. This permits the specification of prior means λ_α and λ_β for σ_α^{-2} and σ_β^{-2} based on previous distribution information, together with prior sample sizes for σ_α^2 and σ_β^2. It should be noted that important sampling (Hsu, Leonard, & Tsui, 1991) and the Gibbs sampler (Gelfand & Smith, 1990) also can be applied to this situation. Comparisons among the above method and other Bayes approaches are needed to provide guidelines for using Bayes methods under IRT.

Conclusion

Many estimation methods have been introduced for estimating item and ability parameters in the context of IRT. There is still a great need for efficient algorithms of the Bayesian approach. In this paper, a procedure is presented for obtaining marginal Bayesian estimates of item parameters with a two-stage hierarchical prior distribution for dichotomously scored IRT models. When the procedure is applied to the simulated data, the item parameter estimates from HB2 are found to agree with estimates from MB.

Appendix

In order to estimate the unknown item parameters, the logarithm of the marginal posterior distribution, $\log p(\boldsymbol{\xi}|\mathbf{y}, \boldsymbol{\tau}, \boldsymbol{\eta}^{(2)}) \propto \log l(\boldsymbol{\xi}|\mathbf{y}, \boldsymbol{\tau}) + \log p(\boldsymbol{\xi}|\boldsymbol{\eta}^{(2)}) = F(\boldsymbol{\xi})$, is maximized by taking partial derivatives with respect to the item parameters and setting them equal to zero. The resulting equations represent the marginal Bayesian estimation equations for item parameters with two-stage priors. As is the case for the maximum likelihood or the nonlinear least squares estimator, we cannot generally solve the estimation equations explicitly for item parameters. Instead, we must solve them iteratively. The Newton-Raphson method and some of its modifications can be used for this purpose (Kennedy & Gentle, 1980). The Newton-Raphson method requires use of both the gradient vector and the Hessian matrix in computations:

$$\hat{\boldsymbol{\xi}}^{(t)} = \hat{\boldsymbol{\xi}}^{(t-1)} - \left\{\mathbf{H}^{(t-1)}\right\}^{-1} \mathbf{f}^{(t-1)}, \tag{22}$$

where

$$\mathbf{f}^{(t-1)} = \left.\frac{\partial F}{\partial \boldsymbol{\xi}}\right|_{\hat{\boldsymbol{\xi}}^{(t-1)}}, \tag{23}$$

$$\mathbf{H}^{(t-1)} = \left.\frac{\partial^2 F}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'}\right|_{\hat{\boldsymbol{\xi}}^{(t-1)}}, \tag{24}$$

and t indexes the iteration. The iteration is repeated until the convergence criterion is met.

Since the dimensionality of all terms in the Newton-Raphson equation is order of $2n$, when the number of items is large, matrices and vectors of considerable size result. These are beyond the capabilities of most digital computers and ways to reduce the dimensionality must be found. We can accomplish this using the EM algorithm (Bock & Aitkin, 1981; Dempster, Laird, & Rubin, 1977). We assume that items are independent, hence, the estimation proceeds one item at a time. The Newton-Raphson equation becomes

$$\hat{\boldsymbol{\xi}}_j^{(t)} = \hat{\boldsymbol{\xi}}_j^{(t-1)} - \left\{\mathbf{H}_j^{(t-1)}\right\}^{-1} \mathbf{f}_j^{(t-1)}. \tag{25}$$

The individual elements which are needed in the Newton-Raphson iteration for the HB2 procedure using the Gaussian quadrature formula are given in Table A.

Insert Table A about here

Note that $\bar{r}_{jk}$ and $\bar{N}_{jk}$ in Table A are defined as

$$\bar{r}_{jk} = \sum_{i=1}^{N} y_{ij} p(X_k|\mathbf{y}_i, \boldsymbol{\xi}, \boldsymbol{\tau}) \tag{26}$$

and

$$\bar{N}_{jk} = \sum_{i=1}^{N} p(X_k|\mathbf{y}_i, \boldsymbol{\xi}, \boldsymbol{\tau}), \tag{27}$$

where

$$p(X_k|\mathbf{y}_i, \boldsymbol{\xi}, \boldsymbol{\tau}) = \frac{\prod_{j=1}^{n} P_j(X_k)^{y_{ij}} Q_j(X_k)^{1-y_{ij}} A(X_k)}{\sum_{k=1}^{q} \prod_{j=1}^{n} P_j(X_k)^{y_{ij}} Q_j(X_k)^{1-y_{ij}} A(X_k)}. \tag{28}$$

Based on provisional item parameter estimates $\hat{\boldsymbol{\xi}}$, obtaining $\bar{r}_{jk}$ and $\bar{N}_{jk}$ is the expectation (E) step of the EM algorithm. The maximization (M) step is to solve the Newton-Raphson equation for each item using obtained provisional $\bar{r}_{jk}$ and $\bar{N}_{jk}$ (Bock & Aitkin, 1981; Bock, Mislevy, & Thissen, 1991). The EM cycles are continued until we obtain a stable set of item parameter estimates.

The EM solution may not provide an estimate of the posterior dispersion matrix. Therefore, to obtain the dispersion matrix, we need to solve the marginal Bayesian estimation equations after obtaining item parameter estimates from the converged EM solution. In this case the $2n \times 2n$ Hessian matrix is

$$\mathbf{H} = -\sum_{i=1}^{N} \frac{1}{p(\mathbf{y}_i|\boldsymbol{\xi}, \boldsymbol{\tau})} \left\{ \frac{\partial p(\mathbf{y}_i|\boldsymbol{\xi}, \boldsymbol{\tau})}{\partial \boldsymbol{\xi}} \frac{\partial p(\mathbf{y}_i|\boldsymbol{\xi}, \boldsymbol{\tau})}{\partial \boldsymbol{\xi}'} \right\} + \frac{\partial^2 \log p(\boldsymbol{\xi}|\boldsymbol{\eta}^{(2)})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'}. \tag{29}$$

The summation in the Hessian matrix, however, involves all examinees and may not be practical to use. When we reformulate the response matrix into distinctive response patterns and the corresponding frequencies, the Hessian matrix may become feasible to calculate. For practical purpose, we can further approximate the Hessian matrix with the use of the empirical information matrix (Bock, Mislevy, & Thissen, 1991). We may need only one or two Newton-Raphson iterations to improve almost converged item parameter estimates of the EM solution.

References

Baker, F. B. (1982). *GENIRV: A program to generate item response vectors.* Unpublished manuscript. Madison: University of Wisconsin, Laboratory of Experimental Design.

Baker, F. B. (1987). Methodology Review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement, 11*, 111–141.

Baker, F. B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17*, 20.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147–162.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer-Verlag.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6*, 258–276.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Applications of an EM algorithm. *Psychometrika, 46*, 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179–197.

Bock, R. D., Mislevy, R. J., & Thissen, D. (1991). *Item response theory.* Unpublished manuscript.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 972–985.

Good, I. J. (1980). Some history of the hierarchical Bayes methodology (with discussion). In J. W. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 489–519). Valencia, Spain: University Press.

Good, I. J. (1983). The robustness of a hierarchical model for multinomials and contingency tables. In G. E. P. Box, T. Leonard, & C. F. Wu (Eds.), *Scientific inference, data analysis, and robustness* (pp. 191–211). New York: Academic Press.

Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement, 15*, 375–389.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variance on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279–291.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and EM algorithm: A didactic. *Journal of Educational Statistics, 13*, 243–271.

Hsu, J. S. J., Leonard, T., & Tsui, K.-W. (1991). Statistical inference for multiple choice tests. *Psychometrika, 56*, 327–348.

Kennedy, W. J., Jr., & Gentle, J. E. (1980). *Statistical computing*. New York: Marcel Dekker.

Kim, S.-H., Cohen, A. S., Baker, F. B., Subkoviak, M. J., & Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika, 59*, 405–421.

Kim, S.-H. (1994, April). *Hierarchical Bayes estimation of item parameters*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Leonard, T. (1982). Comment on the paper by Lejiune and Faukenberry. *Journal of the American Statistical Association*, *77*, 657–658.

Leonard, T., Hsu, J. S. J., & Tsui, K. W. (1989). Bayesian marginal inference. *Journal of the American Statistical Association*, *84*, 1051–1058.

Leonard, T., & Novick, M. R. (1985). *Bayesian inference and diagnostics for the three parameter logistic model* (ONR Technical Report 85-5). Iowa City, IA: The University of Iowa, CADA Research Group. (ERIC Document Reproduction Service No. ED 261 068)

Lindley, D. V. (1971). The estimation of many parameters. In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of statistical inference* (pp. 435–455). Toronto: Holt, Rinehart & Winston of Canada.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, *34*, 1–41.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*, 157–162.

Mendenhall, W., Scheaffer, R. L., & Wackerly, D. D. (1981). *Mathematical statistics with application*. Boston, MA: Duxbury Press.

Mislevy, R. J. (1986). Bayes model estimation in item response models. *Psychometrika*, *51*, 177–195.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57–75.

O'Hagan, A. (1976). On posterior joint and marginal modes. *Biometrika*, *63*, 329–333.

Rudin, W. (1976). *Principles of mathematical analysis* (3rd ed.). New York: McGraw-Hill.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliff, NJ: Prentice-Hall.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, 7, 175–191.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*, 349–364.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, *51*, 581–601.

Thissen, D., & Wainer, H. (1990). Confidence envelopes for item response theory. *Journal of Educational Statistics*, *15*, 113–128.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.

Tsutakawa, R. K. (1984). Estimation of two-parameter logistic item response curves. *Journal of Educational Statistics*, *9*, 263–276.

Tsutakawa, R. K. (1992). Prior distribution for item response curves. *British Journal of Mathematical and Statistical Psychology*, *45*, 51–74.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, *51*, 251–267.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, *52*, 275–291.

Table 1
Root Mean Square Differences (RMSD) and Correlation of Item Discrimination Averaged Over Four Replications

			Hierarchical Bayesian-2			Marginal Bayesian			
	Sample	Item	Prior-α_L	Prior-α_T	Prior-$\alpha\beta_T$	Prior-α_L	Prior-α_T	Prior-$\alpha\beta_T$	ML
RMSD	100	15	.225	.251	.246	.255	.227	.238	.412
	100	45	.233	.276	.276	.255	.231	.238	.348
	300	15	.192	.186	.185	.205	.183	.186	.254
	300	45	.161	.159	.160	.181	.159	.161	.216
Correlation	100	15	.673	.673	.691	.667	.671	.644	.657
	100	45	.688	.686	.693	.679	.682	.671	.676
	300	15	.820	.824	.823	.819	.823	.818	.815
	300	45	.864	.866	.866	.863	.865	.863	.860

Table 2

Root Mean Square Differences (RMSD) and Correlation of Item Difficulty Averaged Over Four Replications

			Hierarchical Bayesian-2			Marginal Bayesian			
	Sample	Item	Prior-α_L	Prior-α_T	Prior-$\alpha\beta_T$	Prior-α_L	Prior-α_T	Prior-$\alpha\beta_T$	ML
RMSD	100	15	.309	.315	.297	.315	.316	.308	.334
	100	45	.284	.290	.269	.298	.287	.277	.352
	300	15	.164	.159	.151	.174	.163	.161	.207
	300	45	.187	.184	.177	.197	.186	.184	.224
Correlation	100	15	.951	.956	.958	.955	.955	.955	.950
	100	45	.963	.962	.964	.958	.962	.963	.942
	300	15	.988	.989	.990	.987	.989	.989	.981
	300	45	.983	.983	.984	.981	.983	.983	.975

Table 3
Mean Euclidean Distances (MED) Averaged Over Four Replications

Sample	Item	Hierarchical Bayesian-2			Marginal Bayesian			ML
		Prior-α_L	Prior-α_T	Prior-$\alpha\beta_T$	Prior-α_L	Prior-α_T	Prior-$\alpha\beta_T$	
100	15	.335	.355	.343	.359	.332	.330	.451
100	45	.320	.342	.331	.344	.322	.317	.423
300	15	.222	.212	.205	.234	.212	.211	.268
300	45	.214	.211	.209	.230	.213	.212	.262

Table A

First and Second Derivatives of the Log Posterior Distribution for HB2

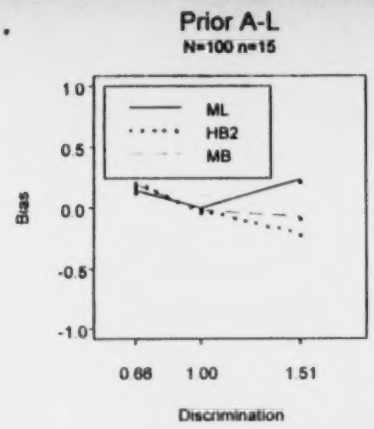
Parameter	Contribution	First Derivative	Second Derivative
α_j	Likelihood	$\exp(\alpha_j)\sum_{k=1}^{q}(X_k-\beta_j)\{\bar{r}_{jk}-\bar{N}_{jk}P_j(X_k)\}$	$-\exp(2\alpha_j)\sum_{k=1}^{q}(X_k-\beta_j)^2P_j(X_k)Q_j(X_k)\bar{N}_{jk}$
α_j	Prior	$-\frac{1}{s_\alpha^2}(\alpha_j-\bar{\alpha})$	$-\frac{1}{s_\alpha^4}\left\{\frac{s_\alpha^2(n-1)}{n}-\frac{2(\alpha_j-\bar{\alpha})^2}{n+\nu_\alpha-1}\right\}$
β_j	Likelihood	$-\exp(\alpha_j)\sum_{k=1}^{q}\{\bar{r}_{jk}-\bar{N}_{jk}P_j(X_k)\}$	$-\exp(2\alpha_j)\sum_{k=1}^{q}P_j(X_k)Q_j(X_k)\bar{N}_{jk}$
β_j	Prior	$-\frac{1}{s_\beta^2}(\beta_j-\bar{\beta})$	$-\frac{1}{s_\beta^4}\left\{\frac{s_\beta^2(n-1)}{n}-\frac{2(\beta_j-\bar{\beta})^2}{n+\nu_\beta-1}\right\}$
$\alpha_j\beta_j$	Likelihood		$\exp(2\alpha_j)\sum_{k=1}^{q}(X_k-\beta_j)P_j(X_k)Q_j(X_k)\bar{N}_{jk}$

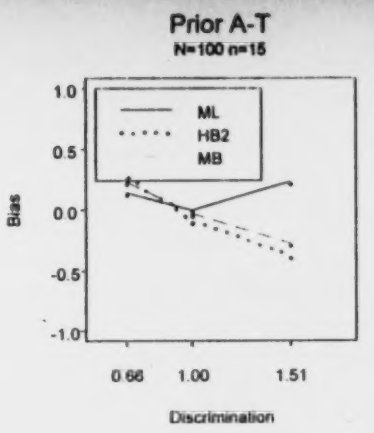
where $s_\alpha^2=\dfrac{\sum_{j=1}^{n}(\alpha_j-\bar{\alpha})^2+\nu_\alpha\lambda_\alpha}{n+\nu_\alpha-1}$, $\bar{\alpha}=n^{-1}\sum_{j=1}^{n}\alpha_j$, $s_\beta^2=\dfrac{\sum_{j=1}^{n}(\beta_j-\bar{\beta})^2+\nu_\beta\lambda_\beta}{n+\nu_\beta-1}$, and $\bar{\beta}=n^{-1}\sum_{j=1}^{n}\beta_j$.

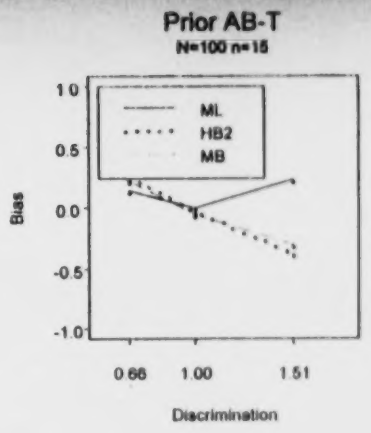
Figure Captions

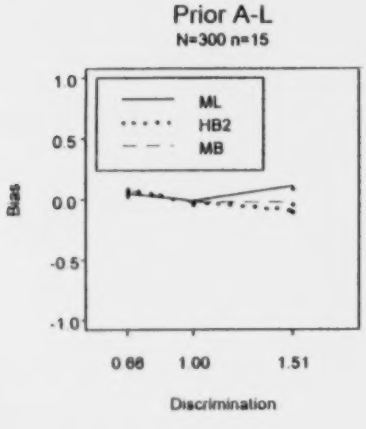
Figure 1. Bias plots for item discrimination.

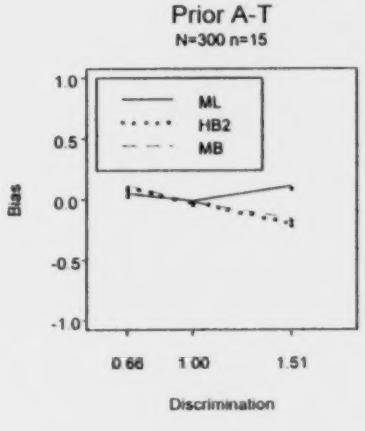
Figure 2. Bias plots for item difficulty.

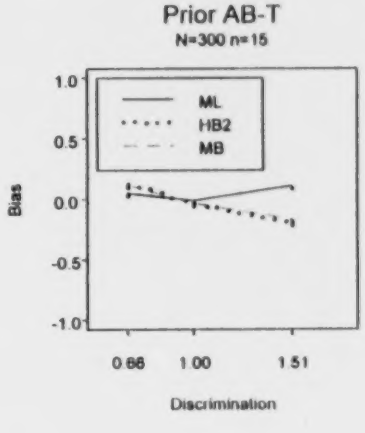

Prior A-L
N=100 n=15
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.66
1.00
1.51
Discrimination

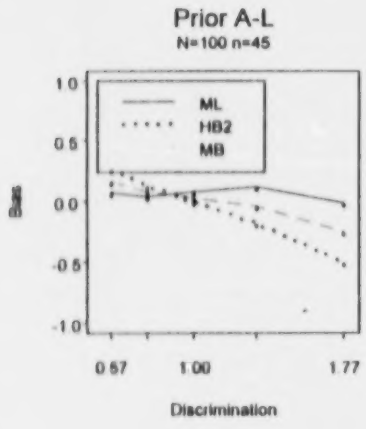

Prior A-T
N=100 n=15
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.66
1.00
1.51
Discrimination

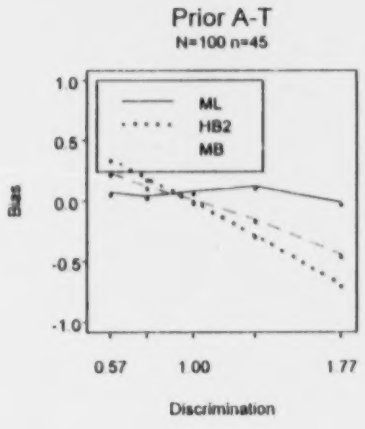

Prior AB-T
N=100 n=15
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.66
1.00
1.51
Discrimination

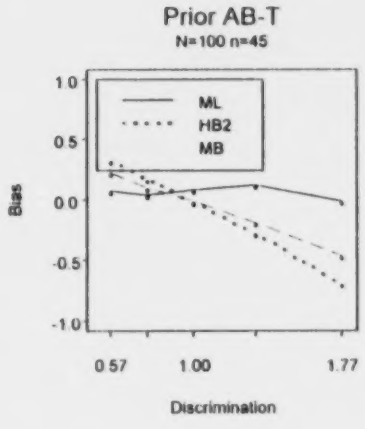

Prior A-L
N=300 n=15
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.66
1.00
1.51
Discrimination

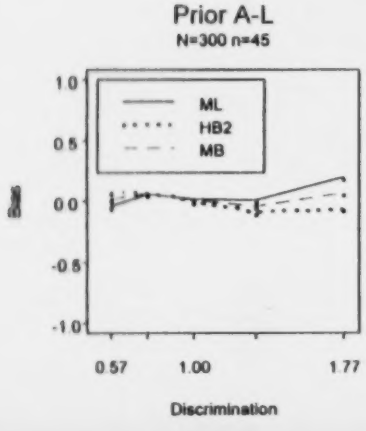

Prior A-T
N=300 n=15
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.66
1.00
1.51
Discrimination

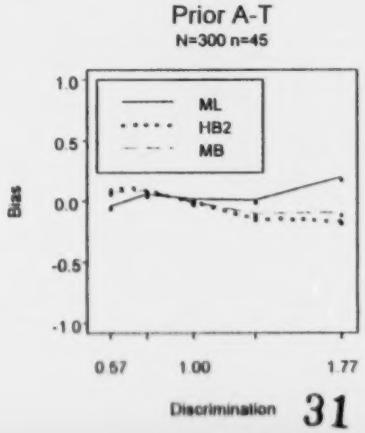

Prior AB-T
N=300 n=15
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.66
1.00
1.51
Discrimination

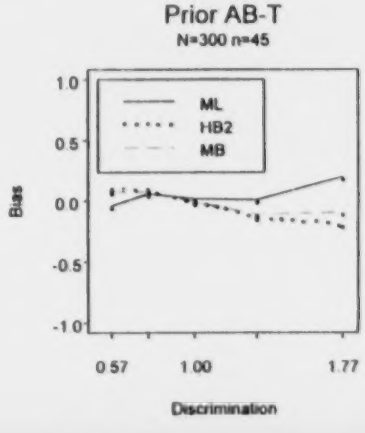

Prior A-L
N=100 n=45
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.57
1.00
1.77
Discrimination

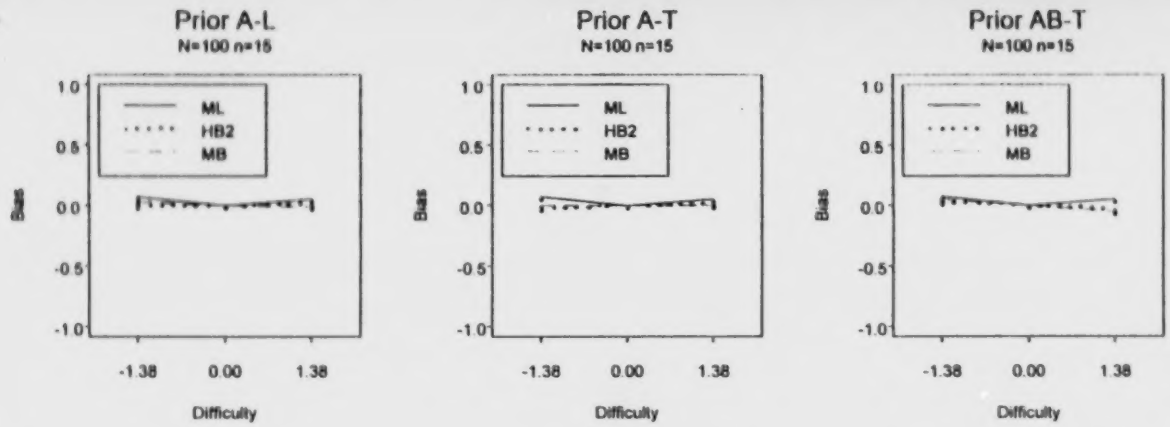

Prior A-T
N=100 n=45
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.57
1.00
1.77
Discrimination

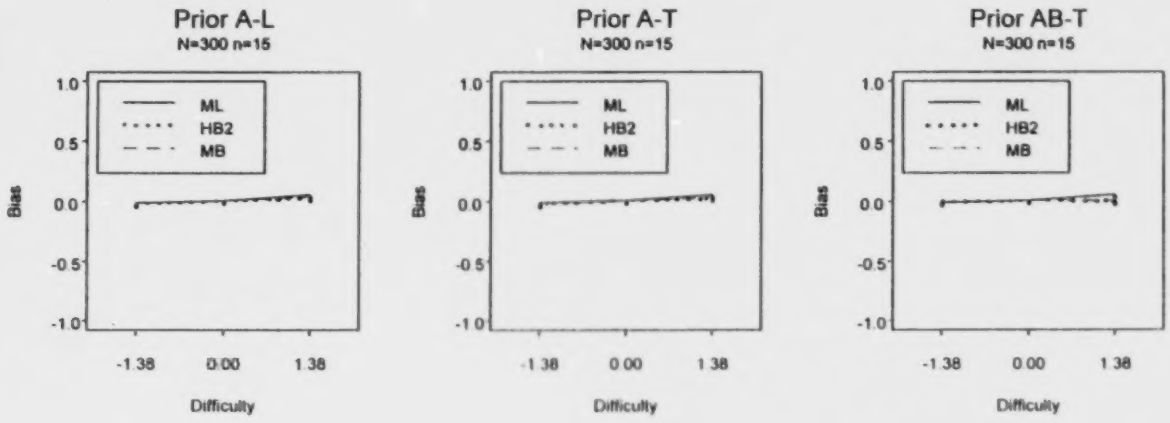

Prior AB-T
N=100 n=45
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.57
1.00
1.77
Discrimination

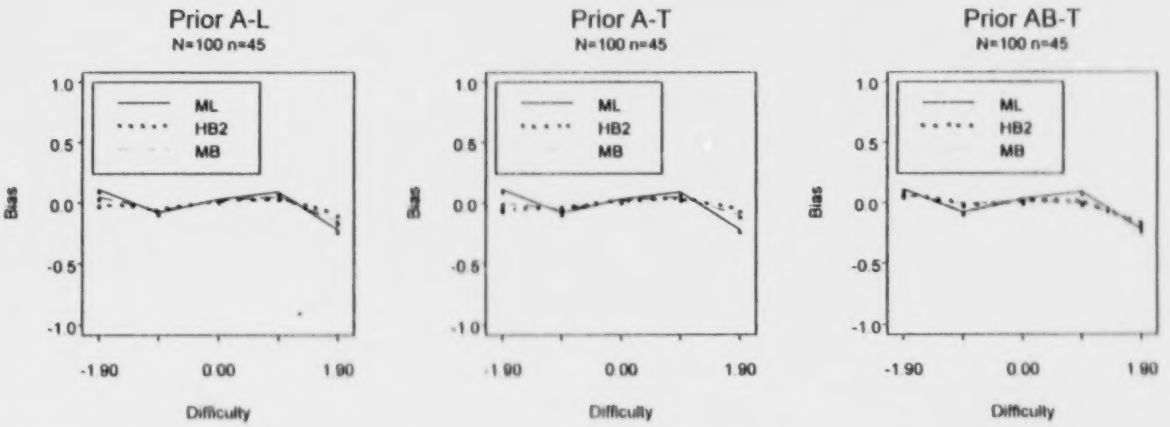

Prior A-L
N=300 n=45
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.57
1.00
1.77
Discrimination

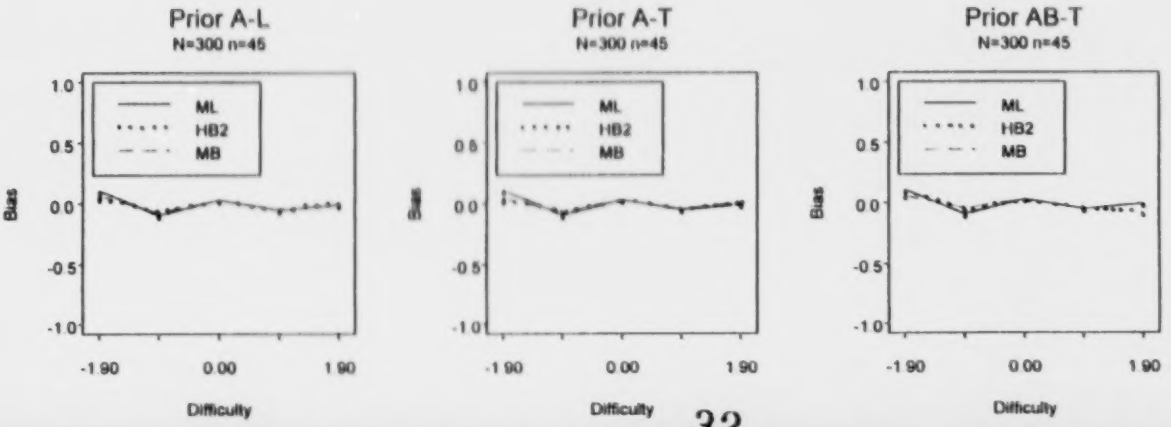

Prior A-T
N=300 n=45
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.57
1.00
1.77
Discrimination


Prior AB-T
N=300 n=45
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
0.57
1.00
1.77
Discrimination


Prior A-L
N=100 n=15
Prior A-T
N=100 n=15
Prior AB-T
N=100 n=15
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
-1.38
0.00
1.38
Difficulty


Prior A-L
N=300 n=15
Prior A-T
N=300 n=15
Prior AB-T
N=300 n=15
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
-1.38
0.00
1.38
Difficulty


Prior A-L
N=100 n=45
Prior A-T
N=100 n=45
Prior AB-T
N=100 n=45
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
-1.90
0.00
1.90
Difficulty


Prior A-L
N=300 n=45
Prior A-T
N=300 n=45
Prior AB-T
N=300 n=45
ML
HB2
MB
Bias
1.0
0.5
0.0
-0.5
-1.0
-1.90
0.00
1.90
Difficulty

Acknowledgments

The author thanks Chris DiStefano for her helpful comments on an earlier draft of the paper.

Author's Address

Send all correspondence to Seock-Ho Kim, The University of Georgia, 325 Aderhold Hall, Athens GA 30602-7143. Internet: skim@coe.uga.edu